

# Dysarthric Speech Database for Universal Access Research

*Heejin Kim<sup>1</sup>, Mark Hasegawa-Johnson<sup>2</sup>, Adrienne Perlman<sup>3</sup>,  
Jon Gunderson<sup>4</sup>, Thomas Huang<sup>2</sup>, Kenneth Watkin<sup>3</sup>, Simone Frame<sup>3</sup>*

<sup>1</sup>Beckman Institute, University of Illinois, Urbana, USA

<sup>2</sup>Department of Electrical and Computer Engineering, University of Illinois, Urbana, USA

<sup>3</sup>Department of Speech and Hearing Science, University of Illinois, Urbana, USA

<sup>4</sup>Division of Disability Resources and Education Services, University of Illinois, Urbana, USA

{hkim17, jhasegaw, aperlman, jongund, t-huang1, watkin, sframe}@uiuc.edu

## Abstract

This paper describes a database of dysarthric speech produced by 19 speakers with cerebral palsy. Speech materials consist of 765 isolated words per speaker: 300 distinct uncommon words and 3 repetitions of digits, computer commands, radio alphabet and common words. Data is recorded through an 8-microphone array and one digital video camera. Our database provides a fundamental resource for automatic speech recognition development for people with neuromotor disability. Research on articulation errors in dysarthria will benefit clinical treatments and contribute to our knowledge of neuromotor mechanisms in speech production. Data files are available via secure ftp upon request.

**Index Terms:** speech recognition, dysarthria, cerebral palsy

## 1. Introduction

A major goal we aim to achieve in the research of automatic speech recognition (ASR) is to develop assistive technologies for people with motor disabilities. ASR development has been successful, reaching 95%-99% accuracy [1], and has provided a useful human-computer interface especially for people who have difficulties in typing with a keyboard. However, individuals with a neuromotor disorder such as cerebral palsy, traumatic brain injury, amyotrophic lateral sclerosis and Parkinson's disease have not been able to utilize the benefit of these advances mainly because their symptoms include motor speech disorder, i.e. dysarthria. Dysarthria is characterized by imprecise articulation of phonemes and monotonic or excessive variation of loudness and pitch [2, 3]. Although dysarthria can differ notably from normal speech due to imprecise articulation, the articulation errors are generally not random, unlike, for example, apraxia. In fact, previous studies show that most articulation errors in dysarthria can be described in terms of a small number of substitution error types [4, 5, 6]. Kent et al. [4], for example, suggest that most articulation errors in dysarthric speech are primarily errors in the production of one distinctive feature. When articulation errors occur in a consistent manner and, as a result, they are predictable, there exists the advantage of using ASR, even for speech that is highly unintelligible for human listeners. Supporting evidence of better performance of ASR than human listeners is reported in Carlson and Bernstein [7].

The necessity of the great amount of training data is a major concern in ASR development especially for dysarthric speakers since speaking can be a tiring task. The only dysarthric speech database currently available is the Whitaker Database [8]. This database consists of 30 repetitions of 46 isolated words (10 dig-

its, 26 alphabet letters, and 10 'control' words) and 35 words from the Grandfather passage produced by each of six individuals with cerebral palsy. This database includes one normal speaker's production of 15 repetitions of the same materials. The Whitaker Database as well as the materials used in most prior research on ASR for dysarthria consists of a small range of words produced by a small number of subjects.

Aiming to develop large-vocabulary dysarthric ASR systems which would allow users to enter unlimited text into a computer, we have constructed a database with a variety of word categories: digits, computer commands, radio alphabet letters, common words selected from the Brown corpus of written English, and uncommon words selected from children's novels digitized by Project Gutenberg (see section 2.2. for more details). We focus particularly on dysarthria associated with cerebral palsy. In addition to audio data, we have recorded video data, because evidence suggests that using visual information aids speech perception [9, 10, 11, 12, 13, 14]. Having data in both modes allows us to test the following, among others, hypotheses: 1) Does an ASR system with both video and audio information perform better than audio-only ASR?, and 2) if so, does the degree of benefit of using visual information vary depending on speakers? We expect our database to be a resource that can be used to improve ASR systems for people with neuromotor disabilities.

Our database is also designed to benefit research in the areas of speech and hearing science and linguistics. As [15] points out, clinicians are becoming more aware of the need to base their treatment decisions on objective data. Phonetic and phonological analysis of dysarthric speech will reveal the characteristics of articulation errors, and can help clinicians maximize the efficiency of clinical treatments. For example, if clinicians know what types of errors occur predominantly for their patients, they can develop specific strategies to modify the errors and improve the level of speech intelligibility. We are in the process of analyzing our database to determine the speaker-specific and the general speaker-independent characteristics of articulation errors in dysarthria, and we encourage other researchers to do the same. Research in this line would yield better understanding of the neuromotor limitations on articulatory movement in dysarthria. Eventually, it will enhance our knowledge of neuromotor mechanisms that underlie human speech production.

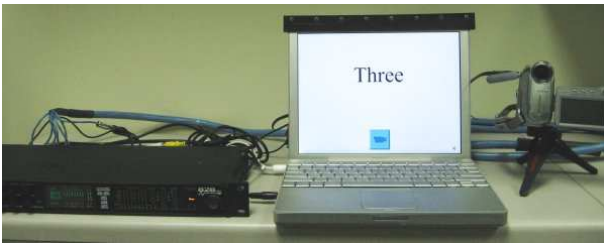


Figure 1: Picture of the equipment: MOTU, the microphone array mounted at the top of a laptop computer, and a video camera.

## 2. Method

### 2.1. Equipment

An eight-microphone array previously designed by our lab [16] was used for recording. Eight microphones were arranged in an array, with 1.5 inches of spacing between adjacent microphones. Each microphone was 6 mm in diameter. The total dimension of the array was roughly 11 inches X 1.5 inches. Microphone preamplifiers were attached to the array. Preamplified audio channels were sent to a multi-channel Firewire audio interface (MOTU 828mkII) through cables. The MOTU recorded eight audio channels at a sampling rate of 48 kHz. One channel of the MOTU was reserved for recording DTMF tones, which were used to segment words and save each word as a separate wav file. The array was mounted at the top of the laptop computer screen, as shown in Figure 1.

In addition to audio recording, one video camera (Canon ZR500) was used to capture the visual features of speech. Audio data recorded at the eighth microphone and DTMF tones were transferred to the video camera. A laptop computer was used to display speech materials and to run the AudioDesk (audio recording software of MOTU). To improve the quality of video recording, we used studio lighting (Lowel Tota-light T1-10) at 750w.

### 2.2. Recording materials and procedures

Most subjects have been recruited through personal contacts established with clients of the Rehabilitation Education Center at the University of Illinois at Urbana-Champaign. Three subjects were recruited in the vicinity of Madison, Wisconsin through a collaboration with the Trace Research and Development Center, University of Wisconsin-Madison. Only subjects who reported a diagnosis of cerebral palsy were recruited.

Recordings (both audio and video) took place while subjects were seated comfortably in front of a laptop computer. Subjects were asked to read an isolated word displayed on a PowerPoint slide on a computer. An experimenter sat beside the subject and advanced the PowerPoint slides after the subject spoke each word. Subjects read three blocks of words. There was always a break between blocks, and subjects were allowed to have a break anytime as needed. Each block contained 255 words: 155 words repeated across blocks, and 100 uncommon words that differed across three blocks. The 155 words included 10 digits ('zero' to 'nine'), 26 radio alphabet letters (e.g., 'Alpha', 'Bravo', 'Charlie'), 19 computer commands (e.g., 'backspace', 'delete', 'enter') and 100 common words (the most common words in the Brown corpus of written English such as 'it, is, you'). The uncommon words (e.g., 'naturalization', 'moonshine', 'exploit') were selected from

children's novels digitized by Project Gutenberg, using a greedy algorithm that maximized token counts of infrequent biphones. In other words, each speaker produced a total of 765 isolated words, including 455 distinct words: three repetitions of 155 words, which allow researchers to train and test whole-word speech recognizers, and 300 distinct uncommon words chosen to maximize phone-sequence diversity. After speech was recorded, the 7 channels of speech for each block were saved as 7 separate wav files. DTMF tones recorded in Channel 1 were then used to automatically segment each wav file into single word files, using Matlab code.

## 3. Description of the database

### 3.1. Speaker

Table 1 summarizes the characteristics of 19 subjects that have been recorded so far. The letter M and F in speaker code specifies a participants' gender. Speech intelligibility (severity of speech disorder) is based on word transcription tasks by human listeners (see section 3.2. for detail). Speakers below the midline were recorded during our preliminary study, reported in Hasegawa-Johnson et al. [17]. Our preliminary research employed the same recording procedures, except that subjects read a total of 541 words: digits three times, and radio alphabet letters, computer commands, words form the 'Grandfather Passage' and words from phonetically balanced sentences (TIMIT sentences [18]) one time each.

Table 1: Summary of speaker information: 'currently being rated' in Speech Intelligibility indicates that intelligibility rate is currently being obtained. (Four speakers below the midline were recorded during our preliminary study).

Speaker	Age	Speech Intelligibility (%)	Dysarthria Diagnosis
M01	>18	very low (10%)	Spastic
M04	>18	very low (2%)	Spastic
M05	21	mid (58%)	Spastic
M06	18	low (39%)	Spastic
M07	58	low (28%)	Spastic
M08	28	currently being rated	Spastic
M09	18	high (86%)	Spastic
M10	21	currently being rated	Mixed
M11	48	mid (62%)	Athetoid
M12	19	currently being rated	Mixed
M13	44	currently being rated	Spastic
M14	40	currently being rated	Spastic
F02	30	low (29%)	Spastic
F03	51	very low (6%)	Spastic
F04	18	mid (62%)	Athetoid
F05	22	high (95%)	Spastic
M01	>18	very low (19%)	Spastic
M02	>18	high (92%)	Spastic
M03	>18	low (29%)	Spastic
F01	>18	low (19%)	Spastic

### 3.2. Intelligibility assessment

Speech intelligibility was assessed to obtain an overall index of severity of dysarthria for each speaker. The purpose of our intelligibility assessment is not to diagnose specific phonetic fea-

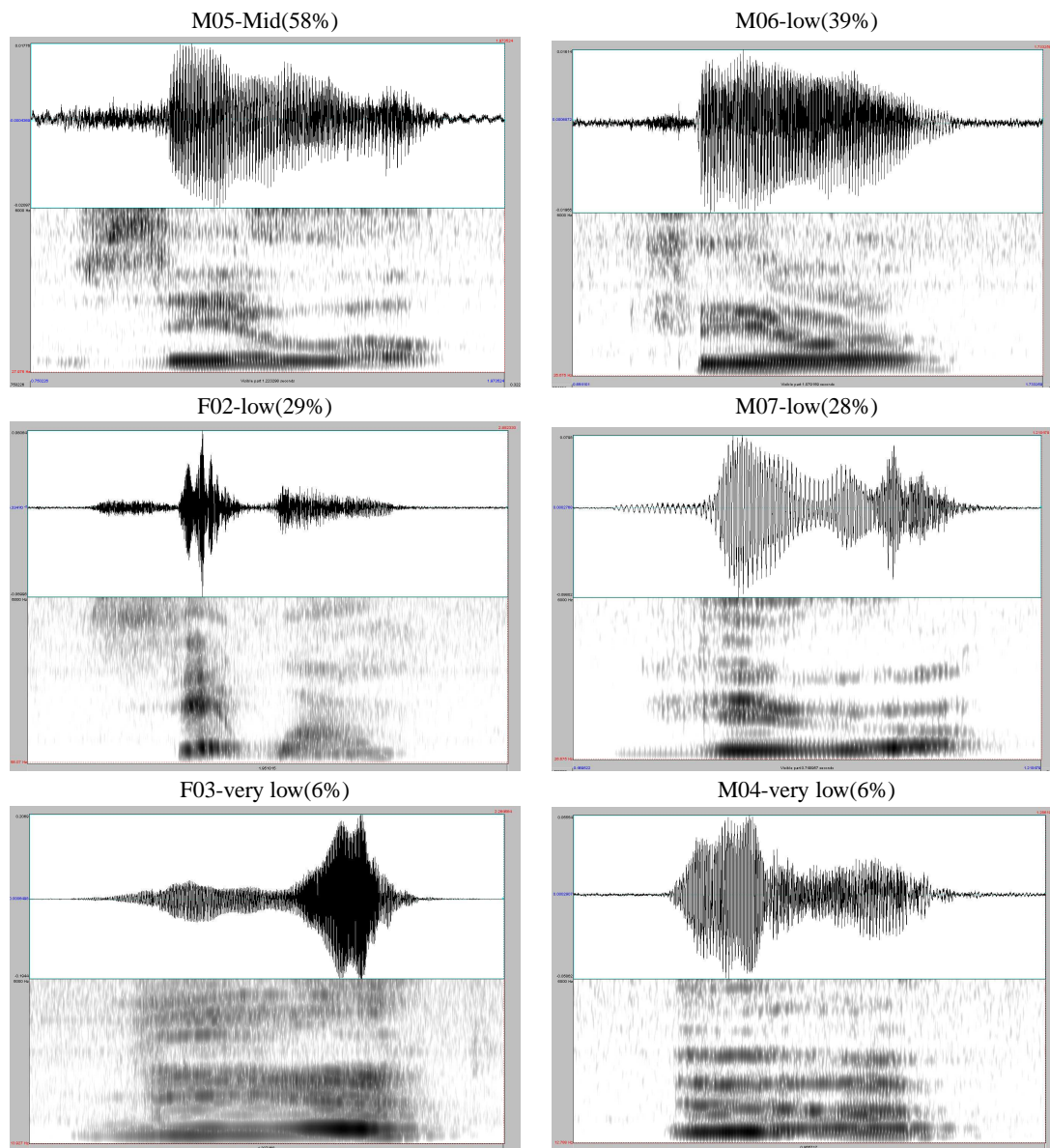


Figure 2: Waveforms and spectrograms of the word 'zero' produced by six different speakers: Speaker code - Intelligibility (%).

tures responsible for reduced intelligibility. Rather we measure the overall intelligibility of speech in order to categorize speakers in terms of intelligibility and to explore any possible relation of articulation error types and ASR architectures to intelligibility levels. For example, we can test if different ASR architectures (e.g. whole-word vs. monophone vs. triphone; Hidden Markov Model (HMM) vs. Support Vector Machines (SVM)) work best for speakers in the high vs. low categories of intelligibility. Hasegawa-Johnson et al. [17] reports that SVM-based recognition is successful for individuals who reduce or delete all consonants, compared to HMM-based recognition.

Two hundred distinct words were selected from the recording of the second block. Words included 10 digits, 25 radio alphabet letters and 19 computer commands. Seventy three words were randomly selected from each category of common and uncommon words. For the purpose of intra-listener reliability assessment, 25 words out of 200 were arbitrarily chosen and

repeated twice in the list. A total of 225 speech files were randomly ordered, with the constraint that repeated words should not be adjacent to each other. The order of words was kept constant across speakers and listeners. Five naive listeners were recruited for each speaker, based on the following criteria: listeners should i) be between 18 - 40 years old, ii) be native speakers of American English, iii) have had no more than incidental experience with persons having speech disorders, iv) have had no training in phonetic transcription, and v) have no identified language disabilities. Listeners were informed that they would be listening to real words spoken by an individual with a speech disorder. They were instructed to provide orthographic transcriptions of each word that they thought the speaker said. Speech files were presented on a web page. They listened to one word at a time with headphones in a quiet room. Listeners were not allowed to use 'not sure' as a response. They were instead asked to write a number (0 to 2) for each word, to indicate

the degree of certainty about their word choice: 0 = not sure at all, 1 = somewhat sure, 2 = completely sure. They were allowed to listen to words as many times as needed. Listeners had three practice items before the actual session started. For each listener's transcription, the percentage of correct responses was calculated. The correct percentage was then averaged across five listeners to obtain each speaker's intelligibility. For the intra-listener reliability assessment, repeated words were examined to see if the transcripts of the two words were identical. For the words marked '2' (i.e. words that listeners were completely sure of), the average agreement rate was 91.64% across listeners. Even for words marked '1' or '0', the two transcriptions were either identical (e.g., the word 'x-ray' was transcribed as 'gray' in both cases) or they were phonologically similar (e.g., 'sink' and 'think' for the word 'sentence'), showing that listeners did not transcribe at random even when words were quite unintelligible. Based on the averaged percent accuracy, each speaker was classified into one of four categories: very low (0-25%), low (26%-50%), mid (51%-75%) and high (76%-100%). For our preliminary data set, intelligibility tests were performed under a similar procedure, except that three listeners transcribed 40 different words from the TIMIT sentences.

### 3.3. Sample data

Figure 2 gives examples of waveforms and spectrograms for the word 'zero' produced by each of six speakers. Their intelligibility rates are displayed next to the Speaker code. Visual inspection of waveforms and spectrograms suggests that different speakers (or different subject groups in terms of intelligibility categories) may exhibit different articulation patterns. For the fricative /z/, the fricative noise is clearly present for Speakers M05, M06 and F02, and less obvious for Speaker M07. For Speakers F03 and M04, /z/ is deleted. In addition, lowering of the third formant that characterizes /r/ is manifested differently depending upon the speaker's intelligibility: considerable lowering of the third formant in M05 and M06, slight lowering in F02 and M07, and almost none for speakers in the 'very low' category (F03, M04). Finally, the formant patterns of the vowels /i/ and /o/ also suggest correlation between articulation patterns and intelligibility. The /i/ and /o/ vowels are well distinguished for the speaker in the 'mid' or some speakers in the 'low' category, as evidenced by a large spacing between the first and second formants for /i/ vs. a much smaller spacing for /o/. As intelligibility decreases, this distinction becomes weaker (e.g. M04's vowel formants rarely change across these two vowels). More detailed phonetic/phonological analysis as well as development of ASR with different recognition architectures is currently in progress in our lab.

## 4. Conclusions

We have described a database of dysarthric speech produced by 19 individuals with cerebral palsy. We are continuing to record new individuals with cerebral palsy. Recording of age-matched normal control subjects is being undertaken. The corpus is available, on request, via secure ftp. We believe that this corpus provides a significant resource for the development of advanced assistive technologies that are beneficial to individuals with neuromotor disorders. Phonetic and phonological analysis of our database will offer an empirical basis for clinical treatment, and further enhance our understanding of neuromotor limitations entailed by articulatory movement in dysarthria and of general mechanisms that underlie speech articulation.

## 5. Acknowledgements

This work has been supported by NSF grant 05-34106 and NIH grant R21-DC008090-A. The authors thank the subjects for their participation and Bowon Lee for building and maintaining a microphone-array. We also thank Mary Sesto and Kate Vanderheiden at University of Wisconsin-Madison for their assistance in recruiting subjects in the Madison area and for allowing us to use the facilities at the Trace Research and Development Center.

## 6. References

- [1] Durham, D., "Key steps to high speech recognition accuracy", Online: <http://www.emicrophones.com>, accessed on 16 September 2007.
- [2] Darley, F., Aronson, A. and Brown, J., "Motor Speech Disorders", Philadelphia: W. B. Saunders Inc., 1975.
- [3] Duffy, J., Motor Speech Disorders, Boston: Elsevier Mosby, 2005.
- [4] Kent, R. D., Weismer, G., Kent, J. F., Vorperian, J. K., and Duffy, J. R., "Acoustic studies of dysarthric speech: Methods, progress, and potential", *Journal of Communication Disorders*, 32:141-186, 1999.
- [5] Platt, L. J., Andrews, G., Young, M., and Quinn, P. T., "Dysarthria of adult cerebral palsy: I. Intelligibility and articulatory impairment", *Journal of Speech and Hearing Research*, 23(1):28-40, 1980.
- [6] Platt, L. J., Andrews, G., Howie, P. M., "Dysarthria of adult cerebral palsy: II. Phonemic analysis of articulation errors", *Journal of Speech and Hearing Research*, 23(1):41-55, 1980.
- [7] Carlson, G. S. and Bernstein, J., "A voice-input communication aid", National Institute of Neurological and Communicative Disorders and Stroke, SRI International, Menlo Park, California, 1988.
- [8] Deller, J. R., Liu, M. S., Ferrier, L. J., and Robichaud, P., "The Whitaker database of dysarthric (cerebral palsy) speech", *Journal of the Acoustical Society of America*, 93(6):3516-3518, 1993.
- [9] Adjoudani, A. and Benoit, C., "On the integration of auditory and visual parameters in an HMM-based ASR", in D.G. Stork and M.E. Hennecke, [Ed], *Speechreading by Humans and Machines: Models, Systems, and Applications*, 461-471, Springer, New York, 1996.
- [10] Chan, M. T., Zhang, Y., and Huang, T., "Real-time lip tracking and audio-video continuous speech recognition", *IEEE Workshop on Multimedia Signal Processing*, Dec. 1998.
- [11] Hennecke, M. E., Stork, D. G., and Prasad, K. V., "Visionary speech: Looking ahead to practical speechreading systems", in D.G. Stork and M.E. Hennecke, [Ed], *Speechreading by Humans and Machines: Models, Systems, and Applications*, 331-350, Springer, New York, 1996.
- [12] Zhang, Y., "Information fusion for robust audio-visual speech recognition", PhD thesis, University of Illinois at Urbana-Champaign, 2000.
- [13] Chu, S. and Huang, T. S., "Bimodal speech recognition using coupled Hidden Markov Models", *Interspeech Proc.*, Beijing, 2000.
- [14] Chu, S. and Huang, T. S., "Multi-modal sensory fusion with application to audio-visual speech recognition", *Eurospeech Proc.*, 2001.
- [15] McHenry, M., "Review of evidence supporting dysarthria treatment in adults", ASHA convention, Boston, November 15-17, 2007.
- [16] Lee, B., Hasegawa-Johnson, M., Goudeseune, C., Kamdar S., Borys, S., Liu, M. and Huang, T., "AVICAR: Audio-visual speech corpus in a car environment", *Interspeech Proc.*, 2004.
- [17] Hasegawa-Johnson, M., Gunderson, J., Perlman, A., and Huang, T., "HMM-Based and SVM-Based Recognition of the Speech of Talkers with Spastic Dysarthria", *ICASSP*, May 2006.
- [18] Zue, V. W., Seneff, S., and Glass J., "Speech database development at MIT: TIMIT and beyond", *Speech Communication*, 9:351-356, 1990.